Earth's Future

Supporting Information for

Identifying key drivers of wildfires in the continental US using machine learning and game theory interpretation

Sally S.-C. Wang[1], Yun Qian[1], L. Ruby Leung[1], and Yang Zhang[2]

[1] Atmospheric Sciences and Global Change Division, Pacific Northwest National Laboratory, Richland, Washington, USA

[2] Department of Civil and Environmental Engineering, Northeastern University, Boston, Massachusetts, USA

Contents of this file

Introduction

The supporting information of "Identifying key drivers of wildfires in the continental US using machine learning and game theory interpretation" consist descriptions about SVD analyses and 15 figures, 3 tables containing many details useful for understanding of the methods and results discussed in the paper.

Text S1. Inverse distance weighting (IDW)

The IDW method represents the attribute value of a grid-cell as the weighted average of known values within a search radius, with the weights inversely related to the distances between the observed points and the grid location, which can be described by the following equation:

$$Z_j = \frac{\sum_{i=1}^{n} w_{ij} z_i}{\sum_{i=1}^{n} w_{ij}} \text{ (S1)}$$

where Z_j is the attribute value of the grid cell *j*, z_i is the observed value at point *i*, $\boldsymbol{w_{ij}} \propto \frac{\mathbf{1}}{\boldsymbol{d_{ij}^p}}$, d_{ij} is the distance between point *i* and the center of grid cell *j*, and *p* is an exponent, and *n* is the number of observations within a search radius. Here we use a value of 1 for *p* and a search radius of 62.5 km, which retains certain heterogeneity among grid cells and produces good correlations between burned area and predictors.

Text S2. The singular value decomposition (SVD) analysis

For the SVD analysis, we consider three regions that periodically experience large wildfires, including Northern California (NCA), southern Rocky mountain area (SRM), and southeastern US (SEUS), as shown in Figure 1a. For each region, we first spatially average the burned area over all the grids to obtain a daily mean burned area time series representative of that region. Then, we calculate the day-to-day correlations between the burned area with five gridded daily meteorological variables (surface temperature, 2-meter RH, U-wind and V-wind at 850 hPa, and geopotential height at 500 hPa) for all 1° × 1° grid cells for the domain, yielding a correlation map for each meteorological variable (Figure. S1). The correlation matrices are then used to derive the SVD modes representing the synoptic weather systems. Note that the daily burned area and meteorological variables are deseasonalized by subtracting 30-day moving averages to focus on day-to-day variability. For a meteorological variable or burned area X_k in a grid, the deseasonalized burned area or meteorology $\tilde{X}_k$ is calculated as follows:

$$\tilde{X}_k(t) = X_k(t) - \frac{1}{31}\sum_{n=t-15}^{t+15} X_k(n) \text{ (S2)}$$

The correlation matrices are then used to derive the SVD modes representing the synoptic weather systems. The correlations of daily mean burned area over NCA with five meteorological variables in the grid boxes of the domain yield a 41 × 35 × 5 (longitude × latitude × variable) matrix, named *D*. Then

we align the dimension of longitude-latitude into one column and transform the matrix *D* into a two-dimensional matrix *A*, with the dimension of 1435 × 5. Then we use matrix *A* to calculate the SVDs, and the SVDs of *A* can be presented as:

$$A = ULV^T \quad \text{(S3)}$$

where *L* is a diagonal matrix, the columns of *V* represent the variable weights, and the columns of *U* represents the spatial weights of the corresponding SVD mode.

We then calculate the time series of the daily magnitude of the two SVD modes using the meteorological fields. We standardize the time series of each meteorological variable in each grid box and obtain a matrix *P*. The magnitude of each SVD mode for the day *t* is then calculated using the inverse process of SVD, which can be expressed as:

$$S_k = Q_k^T P_t M_k \quad \text{(S4)}$$

, where Q_k is the kth column in the spatial weights matrix *Q*, M_k is the kth column in the variable weights matrix, and S_k is a scalar describing the magnitude of kth SVD mode of the meteorological field for that day.

Lastly, we compute the monthly standard deviation of the daily SVD time series obtained above, representing the month-to-month synoptic fluctuations and instability.

Text S3. The SVD analysis for NCA, SRM, and SEUS

This section will demonstrate the results of correlation analysis, derived SVDs, and the monthly standard deviation of the daily SVD time series for NCA, SRM, and SEUS. Figure S1a shows the correlations between daily mean burned area over NCA with the daily mean RH in grid cells across the western US (25º - 60ºN, 140º-100ºW) during 2000-2017. Negative correlations are found not only over NCA but also the neighboring regions such as Nevada and Arizona. The burned area is positively correlated with temperature over California, with the strongest positive correlation over southern California (Figure S1b). The relationships between NCA burned area and geopotential height also show similar patterns, with the strongest positive correlation over Oregon (Figure S1c). Strong negative correlations between NCA burned area and east-west wind speeds are found across the western US and eastern Pacific Ocean (Figure S1d). The relationships between NCA burned areas with north-south wind speeds are relatively weak, with negative correlations over continental western US (Figure S1e). The correlation analysis suggests that large burned areas are related to a high-pressure system over Pacific Northwest and NCA with hot and dry conditions, bringing northeasterly inland flows to California. The variable weights show the major components of the SVD modes, and the spatial weights indicate the spatial distributions of the SVD modes. Figure S2a and S2c shows the spatial and variable weights of the first SVD mode (SVD1). The SVD1 explains 36% of the

total variance, dominated by the positive anomalies over the NCA region (Figure S2a), corresponding to low RH, high temperature, high pressure, and northeasterly winds (Figure S2c). The SVD2 explains 25% of the total variance, characterized by positive anomalies inland, associated with low RH and northwesterly winds (Figure S2b and S2d). Figure S2e and S2f shows the time series of the monthly standard deviation of the two SVDs and the monthly burned area over NCA. Moderate negative correlations are found between both SVDs and burned area (r = -0.47 for SVD1; r = -0.52 for SVD2), indicating that large burned area is associated with the more stagnant synoptic weather characterized by a high-pressure system over the western US, related to northeasterly inland flows to California.

The above analysis is also conducted for SRM and SEUS. Fig. S3 shows the correlations between daily mean burned area over SRM with the daily mean meteorology in 1° × 1° grid cells across the southwestern US (29° -42°N, 119° -100°W) over 2000-2017. The correlation between SRM burned area and RH is negative across the domain, and the strongest correlation is over the SRM area (Figure. S3a). The positive correlations between burned area and temperature are observed across the whole domain, indicating the influence of regional meteorology on burned areas over this region (Figure. S3b). Similar to temperature, positive correlations between burned area and geopotential height are found across the whole domain, with the strongest correlation over the border of Texas and New Mexico (Figure. S3c). The correlations between burned area and wind speeds are relatively weak, which may be explained by the complex topography over this region (Figure. S3d-e). We then utilize the correlation matrix to obtain the SVD modes. Figure. S4a-d shows the spatial and variable weights of the two SVD modes. The SVD1 of the SRM area can explain 60% of the total variance, corresponding to high temperature, high pressure, and low RH across the whole domain. The SVD2 explains much less than the SVD1 (only 13%), composed of northeasterly winds and high pressure over the domain. The synoptic circulation patterns associated with the SRM burned area features a high-pressure system with hot and dry conditions over the SRM area (SVD1). When comparing the monthly standard deviation of daily SVDs with the monthly burned area, we find the negative correlations between burned area and the monthly standard deviation of daily SVDs (r = -0.46 and r = -0.42 for SVD1 and SVD2, respectively; Figure. S4e-f). Identical to NCA, the negative correlations can be interpreted as the more stagnant fire-favorable conditions (i.e., less variation

in the synoptic patterns of high-pressure center and hot-dry conditions) tend to cause large burned area.

Fig. S5a shows the correlations between daily mean burned area over SEUS with the daily mean RH in 1° × 1° grid cells across the southeastern US (25°-40°N, 100°-61°W) over 2000-2017. Negative correlations appear across the southeastern US, with the strongest correlation over Georgia and Florida (the selected domain for calculating mean burned area). The relationship between burned area and temperature shows a regional characteristic, where the positive correlations are observed over the west part of the southeastern US, and the negative correlations are over the east part of the southeastern US (Figure. S5b). Similar patterns are shown for the geopotential height (Figure. S5c). Interestingly, the relationships of burned area over the SEUS with the east-west wind speeds show a strong bimodal structure, with negative correlations over the Gulf of Mexico and positive correlations over the north part of the southeastern US (Figure. S5d). The bimodal structure is also observed in the relationship between burned area and north-south wind speeds, with positive correlations over the west of the southeastern US and negative correlations over the east of the southeastern US (Figure. S5e). The correlations between the burned area and weed speeds suggest that the burned area over east of the southeastern US is associated with a regional-scale anticyclonic circulation. In terms of the SVD modes, the spatial weight of SVD1 shows negative anomalies over east of the SEUS, corresponding to the high RH, southerly winds, and high pressure (Figure. S6a and S6c). For the SVD2, it explains 24% of the total variance. The positive anomalies over SEUS correspond to high RH and low pressure (Figure. S6b and S6d). The identified SVDs suggest a relatively wet condition and southerly flows, unfavorable for burned areas in the SEUS. Then we compute the magnitude of SVDs and derive the daily time series of the magnitude of the two SVDs. The correlations between monthly burned area and standard deviation of daily SVDs are relatively weak ($r = 0.17$ and $r = 0.14$ for the SVD1 and SVD2, respectively). The largest correlation with the monthly burned area appears when the monthly standard deviation of SVDs is at a 2-month lag ($r = 0.5$ for SVD1 and $r = 0.39$ for SVD2; Fig. S6e-f). Therefore, we choose the 2-month lag monthly standard deviation of two SVDs of SEUS. The detailed discussions of the SVDs of SEUS are in Section 2.2.2 and Section 5.2.

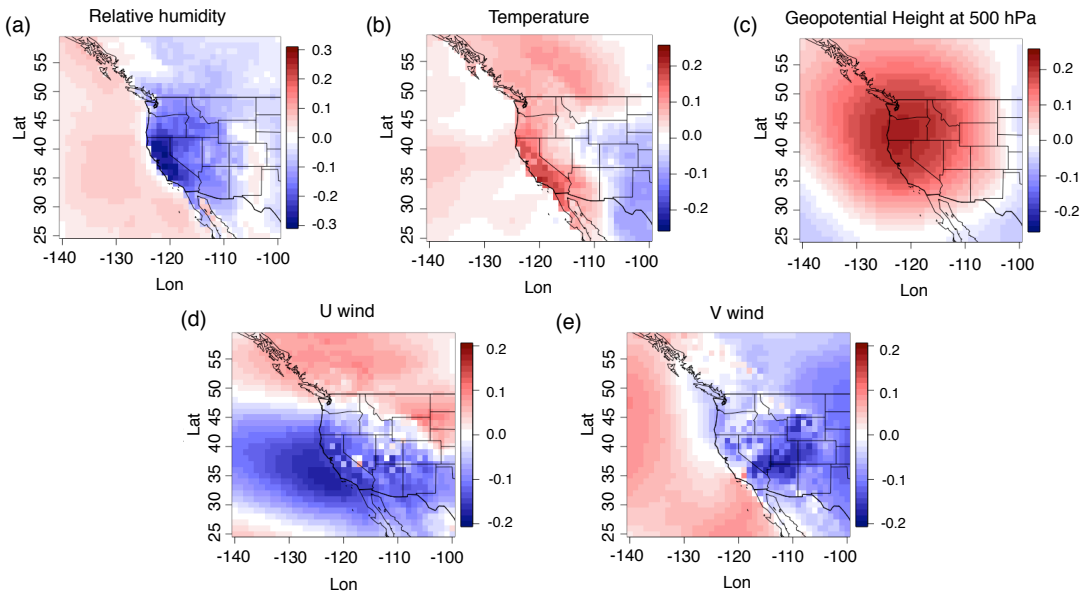

Figure S1. Observed correlations of daily mean burned area over NCA with the meteorology of all grid boxes in the western US from 2000 to 2017. The meteorological variables from NARR include (a) relative humidity at 10 m, (b) surface temperature, (c) geopotential height at 500 hPa, (d) u-wind speed at 850 hPa, and (e) v-wind speed at 850 hPa. All data are deseasonalized by 30-day moving average from the daily values.

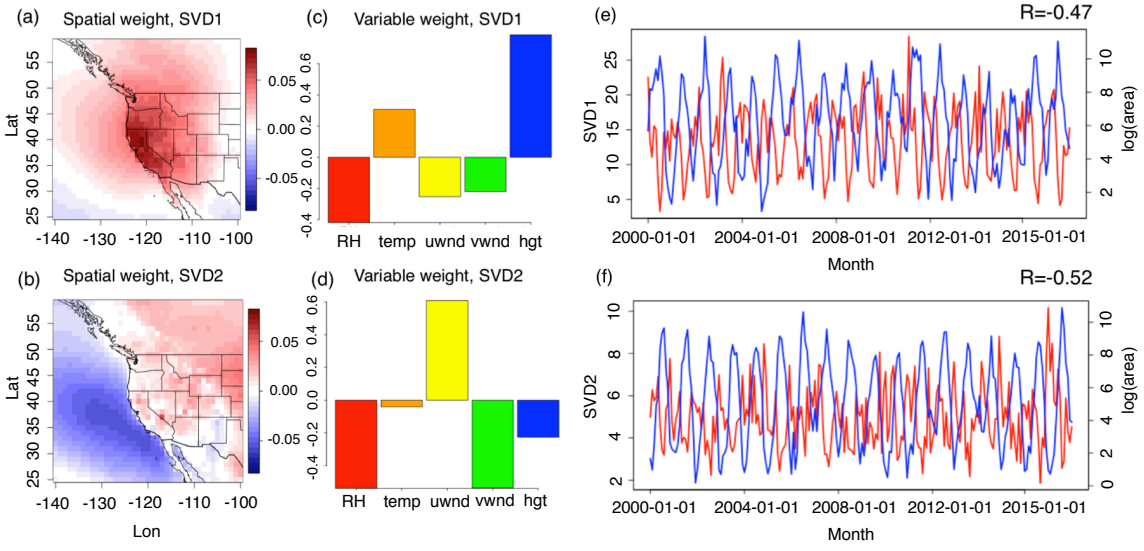

Figure S2. The spatial and variable weights of the first (a, c) and second (b, d) singular value decomposition (SVD) modes describing the spatial correlations between daily mean burned area over NCA and meteorological variables in the grid boxes in the western US from 2000 to 2017. The meteorological variables include temperature, relative humidity, precipitation, u and v wind speed. (e) Time series of monthly standard deviation of daily SVD1 (red) and logarithm of monthly total burned area (blue) for NCA. (f) Time series of monthly standard deviation of daily SVD2 (red) and logarithm of monthly total burned area (blue) for NCA.

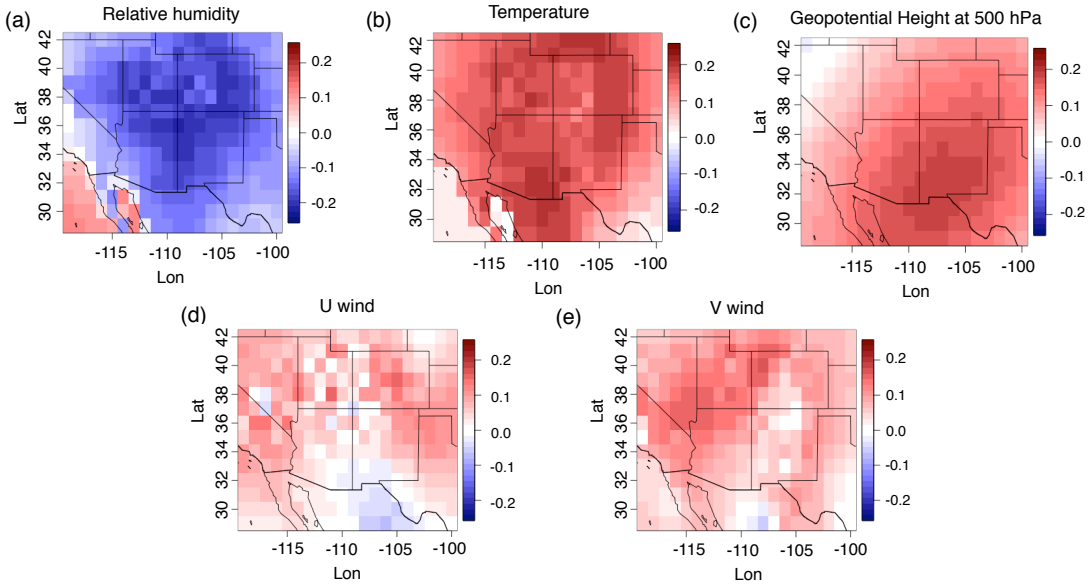

Figure S3. Observed correlations of daily mean burned area over SRM with the meteorology of all grid boxes in the western US from 2000 to 2017. The meteorological variables from NARR include (a) relative humidity at 10 m, (b) surface temperature, (c) geopotential height at 500 hPa, (d) u-wind speed at 850 hPa, and (e) v-wind speed at 850 hPa. All data are deseasonalized by 30-day moving average from the daily values.

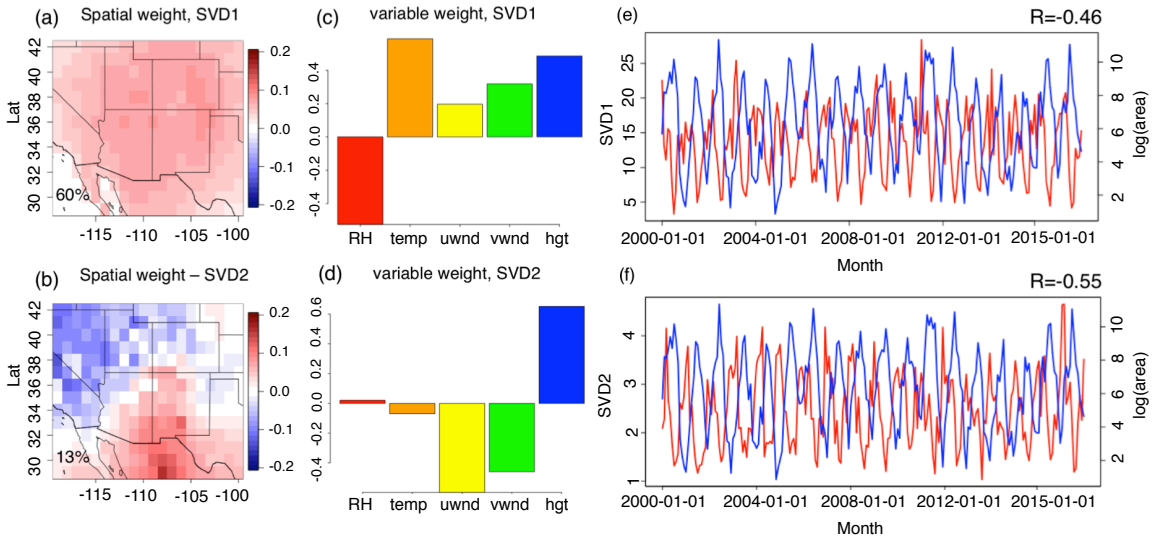

Figure S4. The spatial and variable weights of the first (a, c) and second (b, d) singular value decomposition (SVD) modes describing the spatial correlations between daily mean burned area over SRM and meteorological variables in the grid boxes in the western US from 2000 to 2017. The meteorological variables include temperature, relative humidity, precipitation, u and v wind speed. (e) Time series of monthly standard deviation of daily SVD1 (red) and logarithm of monthly total burned area (blue) for SRM. (f) Time series of monthly standard deviation of daily SVD2 (red) and logarithm of monthly total burned area (blue) for SRM.

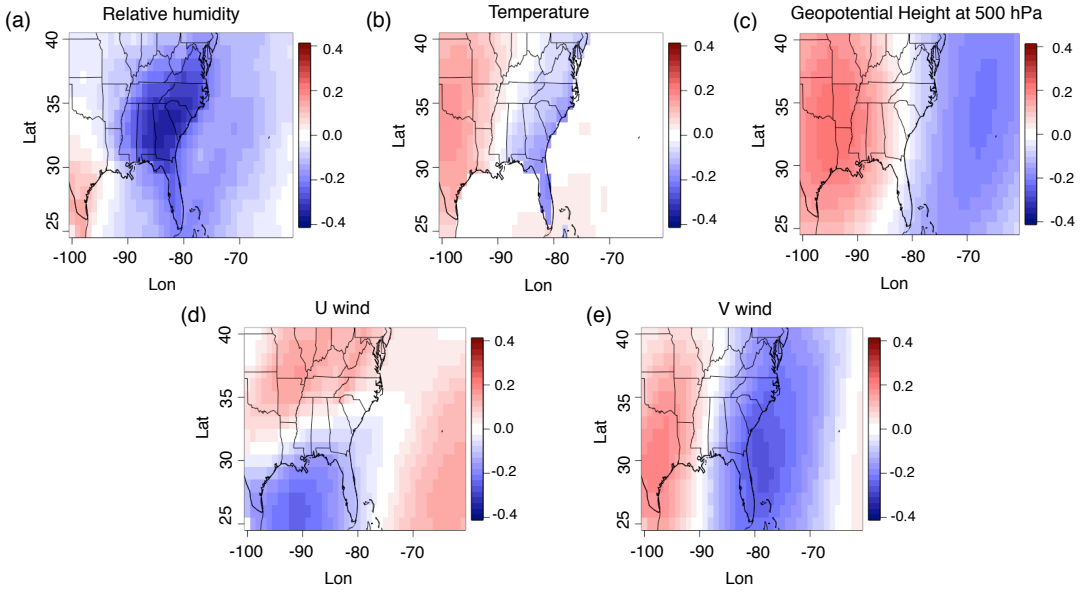

Figure S5. Observed correlations of daily mean burned area over SEUS with the meteorology of all grid boxes in the southeastern US from 2000 to 2017. The meteorological variables from NARR include (a) relative humidity at 10 m, (b) surface temperature, (c) geopotential height at 500 hPa, (d) u-wind speed at 850 hPa, and (e) v-wind speed at 850 hPa. All data are deseasonalized by 30-day moving average from the daily values.

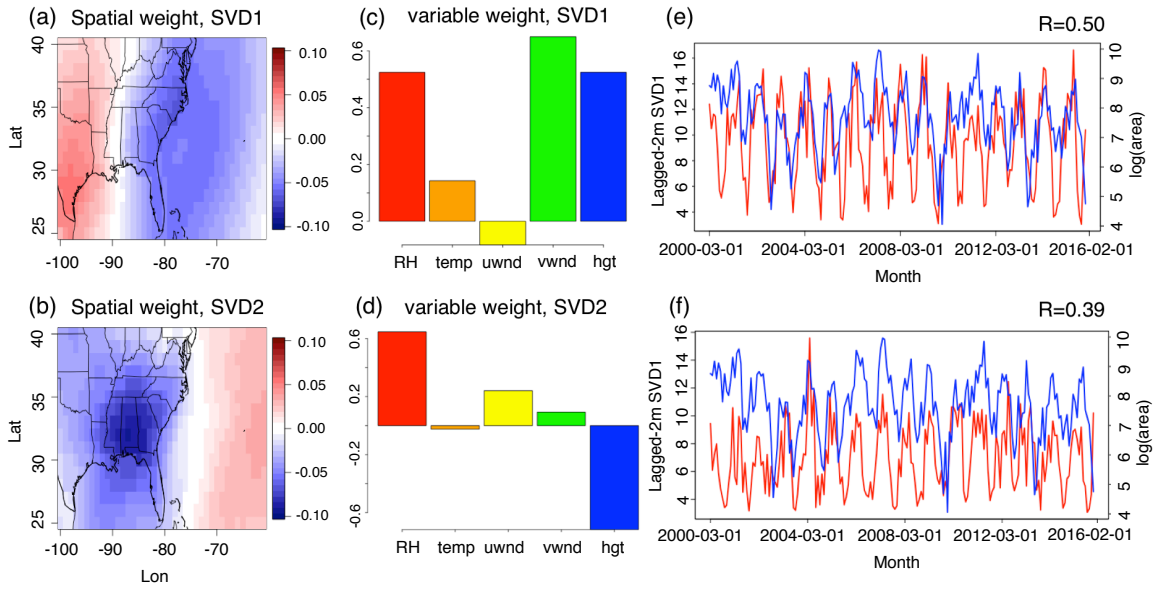

Figure S6. The spatial and variable weights of the first (a, c) and second (b, d) singular value decomposition (SVD) modes describing the spatial correlations between daily mean burned area over SEUS and meteorological variables in the grid boxes in the southeastern US from 2000 to 2017. The meteorological variables include temperature, relative humidity, precipitation, u and v wind speed. (e) Time series of monthly standard deviation of daily SVD1 (red) at a 2-month lag and logarithm of monthly total burned area (blue) for SEUS. (f) Time series of monthly standard deviation of daily SVD2 (red) at a 2-month lag and logarithm of monthly total burned area (blue) for SEUS.

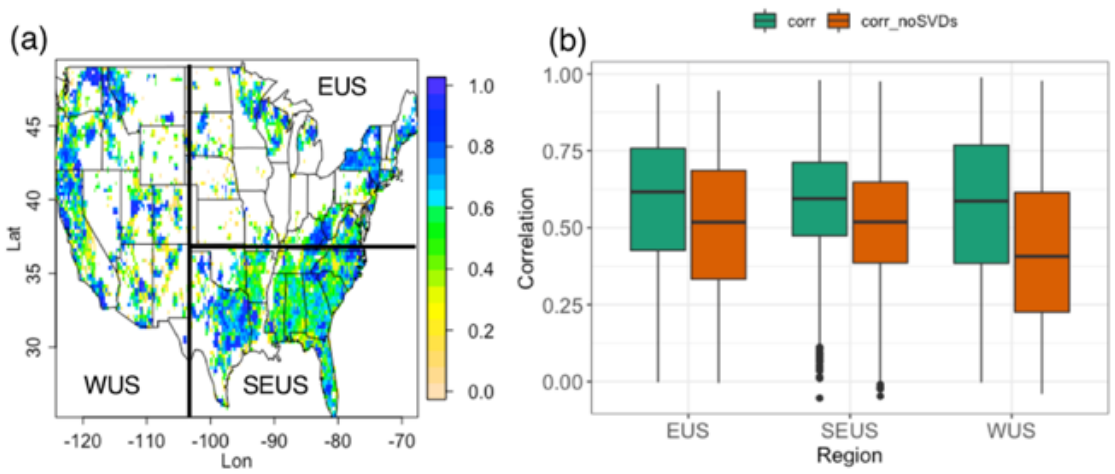

Figure S7. (a) Map of three regions for model performance comparison. (b) Comparison of distributions of correlation between predictions with and without SVD predictors for the three regions.

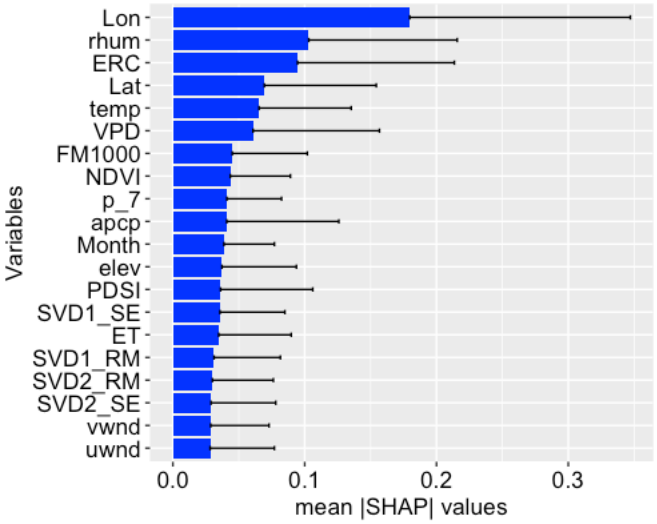

Figure S8. Top 20 variables based on the mean absolute SHAP values with error bars of standard deviation.

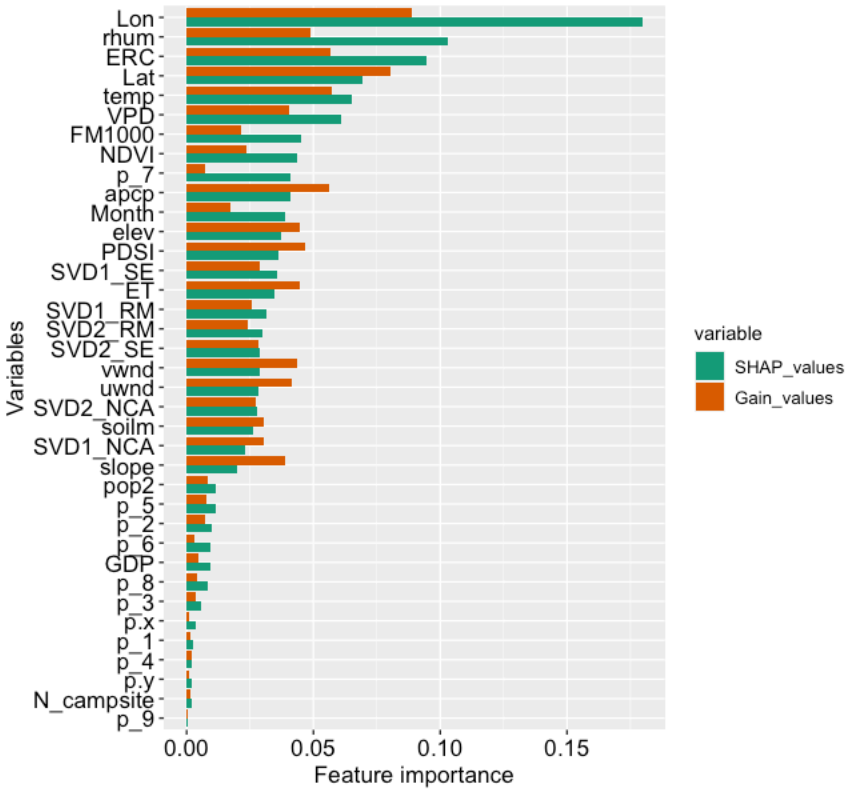

Figure S9. Top 20 variables of the model based on the feature importance determined by the mean absolute SHAP values (green) and gain (orange).

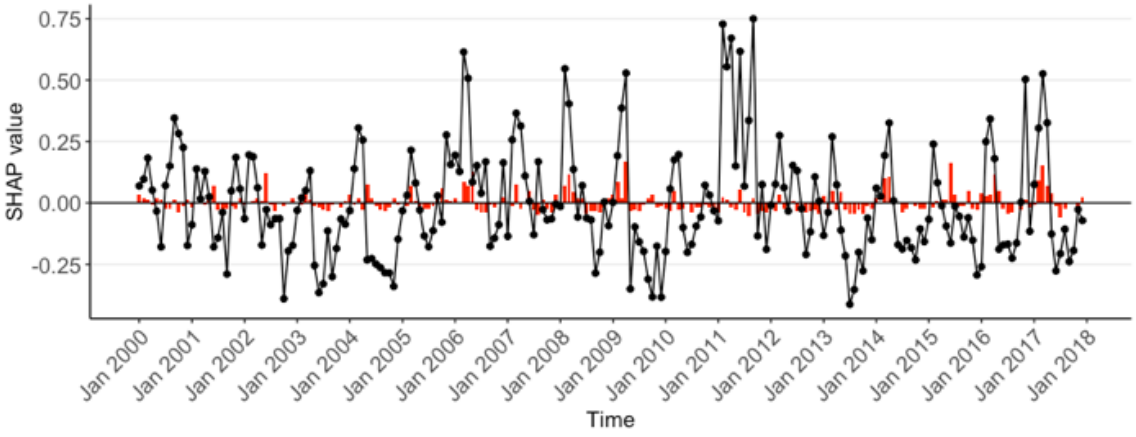

Figure S10. Time series of SHAP values of SVD1_SElag2 (red bar) and predicted normalized burned area (black line).

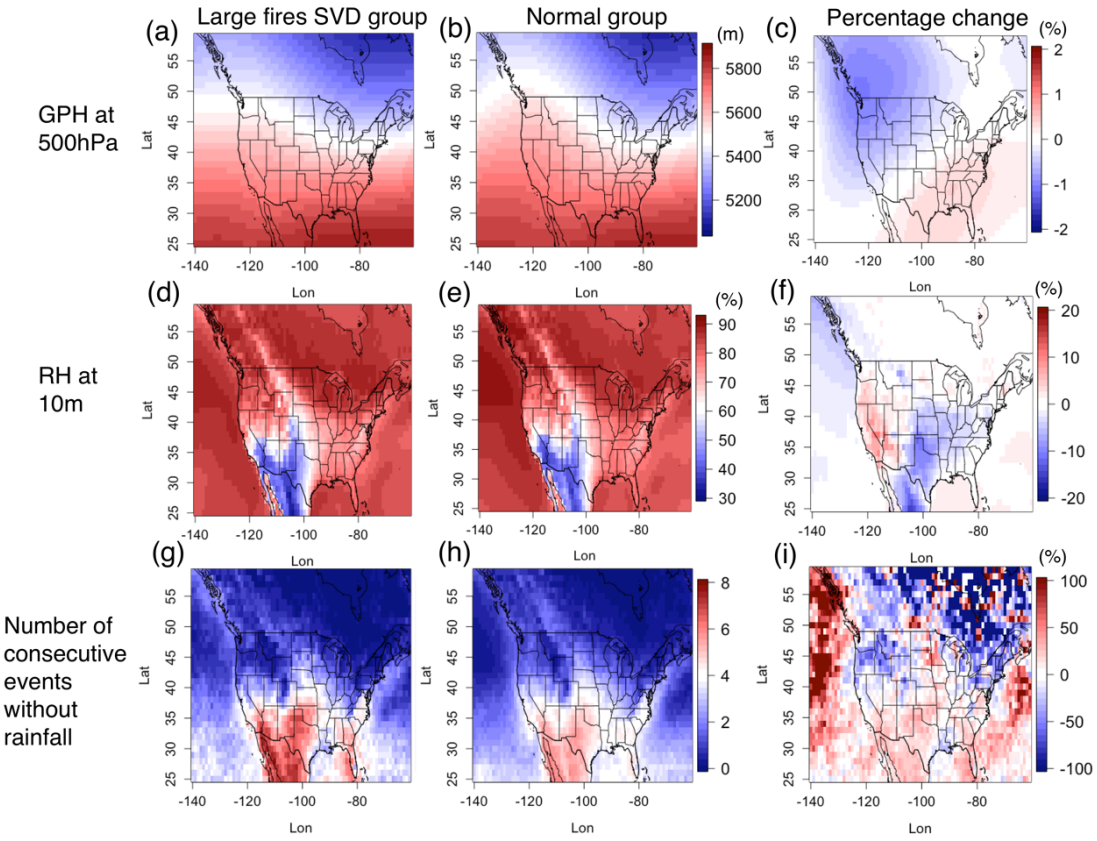

Figure S11. The composite of geopotential height at 500 hPa, RH at 10 m, and number of consecutive days without rainfall for the large fire SVD group and normal group and the difference between them in percentage (large fires SVD group minus normal group divided by normal group).

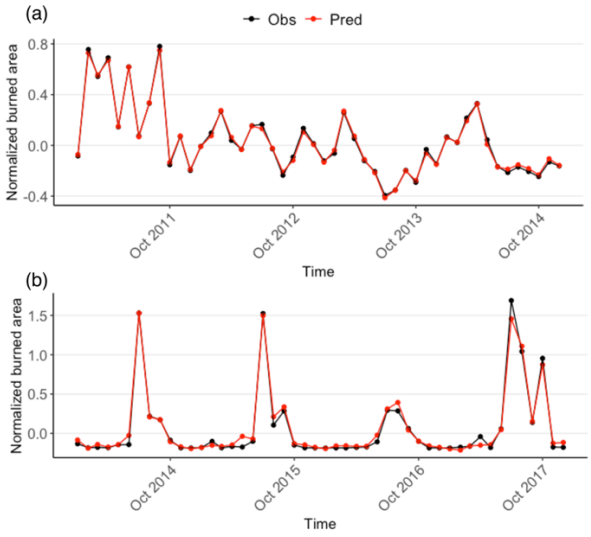

Figure S12. Time series of observed and predicted normalized burned area for (a) SEUS and (b) NCA.

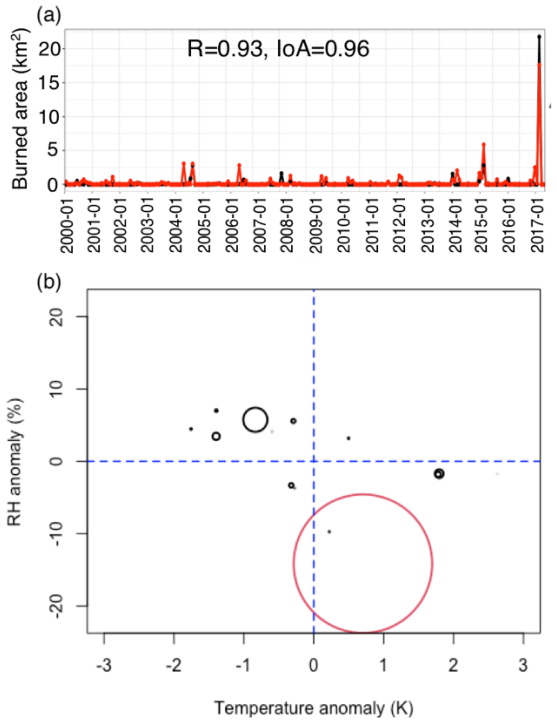

Figure S13. (a) Time series of observed (black) and predicted (red) burned area for a grid at 122.375°W, 38.375°N (Sonoma City, California). (b) Scatter plot of the monthly RH anomaly and temperature anomaly for the same grid during 2000-2017. The size of the circle indicates the burned area. The red circle represents the episodic event in Oct 2017.

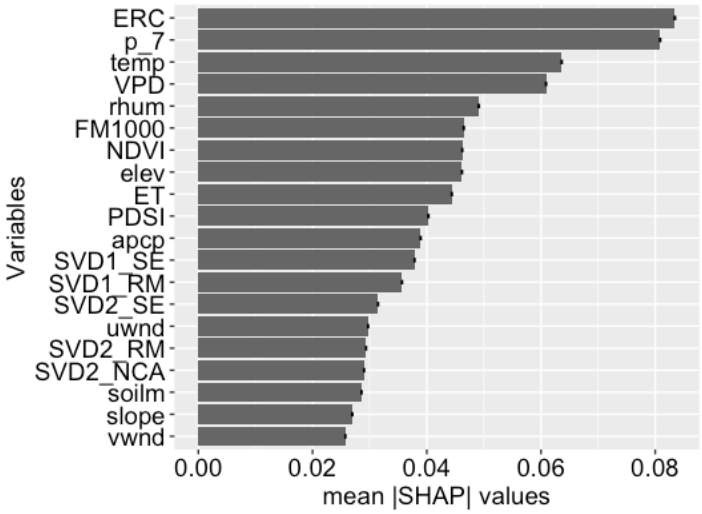

Figure S14. Top 20 variables for the model without variables of longitude, latitude, and month based on the mean absolute SHAP value with error bars of 95% confidence intervals.

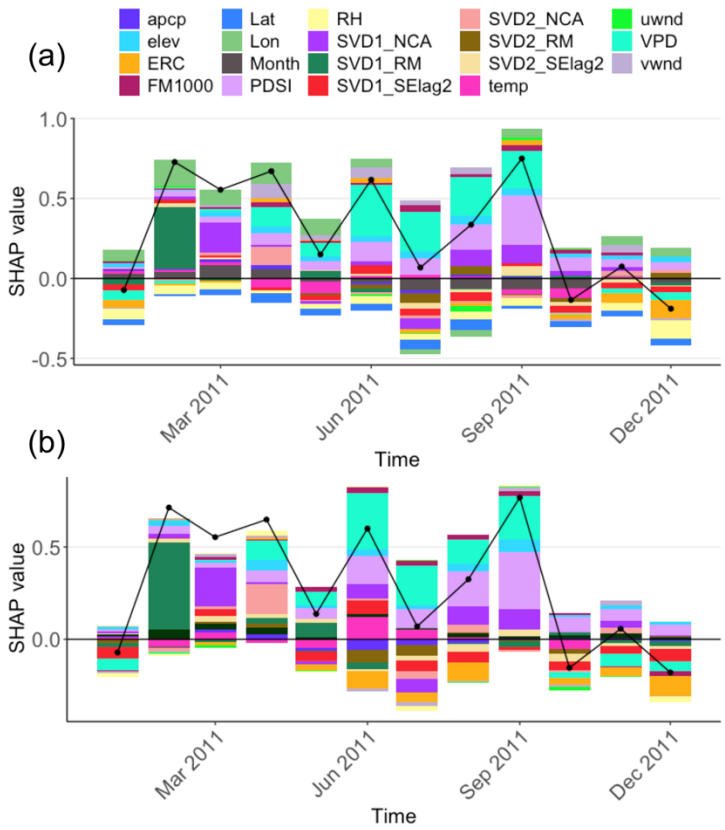

Figure S15. Time series of the average SHAP values (bar) and predicted normalized burned area (line) for the SEUS in 2011 for (a) the original model and (b) the model excluding variables of longitude and latitude.

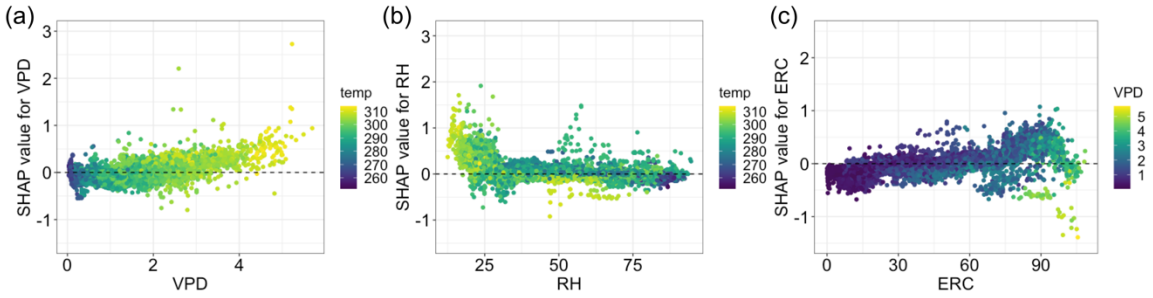

Figure S16. The SHAP dependence plots (a) between SHAP values and VPD, showing the interaction with temperature (color scale); (b) between SHAP values and RH, showing interaction with temperature (color scale); (c) between SHAP values and ERC, showing

interaction with VPD (color scale) for year 2009 and the model without variables of longitude, latitude, and month.

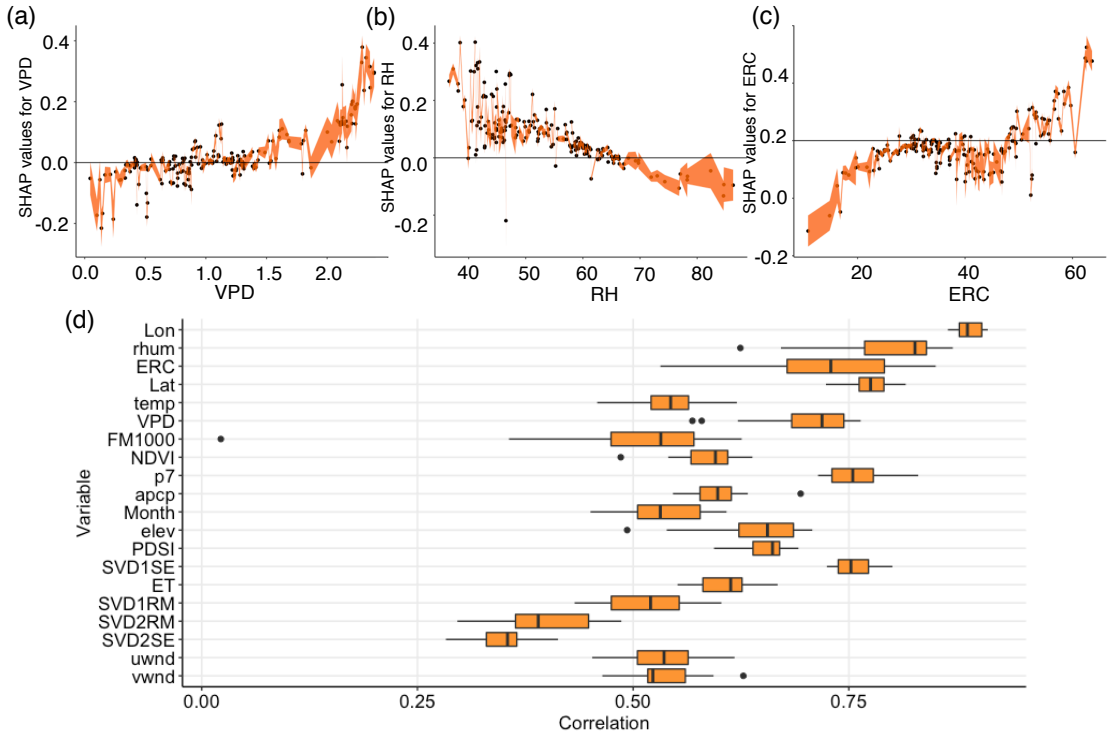

Figure S17. Top: Plots of the SHAP values for the randomly selected 200 sample points in the year 2009 for different predictors: (a) VPD, (b) RH, and (c) ERC. The orange shading indicates the 95% confidence interval calculated from the 20-time bootstrapping models. Bottom: (d) Box plots of correlations between the SHAP values derived from the original and the 20 retrained models for the top 20 variables.

Table S1. Predictor variables used in the fire prediction model

Variables	Abbreviation	Categories	Temporal resolution	Spatial resolution	Data Source	References
Monthly mean surface temperature	temp	Local meteorology	monthly	32 km	North American Reanalysis (NARR)	Mesinger et al. (2006)
Monthly mean relative humidity	RH	Local meteorology	monthly	32 km	North American Reanalysis (NARR)	Mesinger et al. (2006)
Monthly mean of daily precipitation	precip	Local meteorology	monthly	32 km	North American Reanalysis (NARR)	Mesinger et al. (2006)
Monthly mean zonal component of wind speed	U	Local meteorology	monthly	32 km	North American Reanalysis (NARR)	Mesinger et al. (2006)
Monthly mean meridional component of wind speed	V	Local meteorology	monthly	32 km	North American Reanalysis (NARR)	Mesinger et al. (2006)
Monthly mean Palmer Drought Severity Index	PDSI	Local meteorology	8-days	4 km	gridMET	Abatzoglou (2013)
Monthly mean 1000-hour dead fuel moisture	FM1000	Local meteorology	daily	4 km	gridMET	Abatzoglou (2013)
Monthly mean energy release component	ERC	Local meteorology	daily	4 km	gridMET	Abatzoglou (2013)
Monthly mean vapor pressure deficit	VPD	Local meteorology	daily	4 km	gridMET	Abatzoglou (2013)
Monthly standard deviation of daily	SVD1_NCA and SVD2_NCA	Large-scale meteorological patterns	monthly	Regional	North American Reanalysis (NARR)	This study (sect 2.2.2)

SVDs for northern California						
Monthly standard deviation of daily SVDs for southern Rocky Mountain	SVD1_SRM and SVD2_SRM	Large-scale meteorological patterns	monthly	Regional	North American Reanalysis (NARR)	This study (sect 2.2.2)
Monthly standard deviation of daily SVDs for southeastern US (with 2-month lag)	SVD1_SElag2 and SVD2_SElag2	Large-scale meteorological patterns	monthly	Regional	North American Reanalysis (NARR)	This study (sect 2.2.2)
Monthly mean evapotranspiration	ET	Land-surface properties	monthly	0.25°x0.25°	Global Land Data Assimilation System (GLDAS-2)	Rodell et al. (2004)
Monthly mean surface soil moisture	soilm	Land-surface properties	monthly	0.25°x0.25°	Global Land Data Assimilation System (GLDAS-2)	Rodell et al. (2004)
Normalized difference vegetation index	NDVI	Land-surface properties	monthly	0.05°x0.05°	MODerate resolution Imaging Spectroradiometer (MODIS)	Didan (2015)
Land cover percentage	p_	Land-surface properties	Yearly	0.05°x0.05°	MODerate resolution Imaging Spectroradiometer (MODIS); LAI classification scheme	Friedl (2015)
Median Topography (slope and elevation)	Slope and elevation	Land-surface properties	Not change by time	100 km		Amatulli et al. (2018)
Gross domestic product	GDP	Socioeconomic and coordinate variables	Yearly	5 arc		Kummu et al. (2018)

Population density	Pop	Socioeconomic and coordinate variables	Yearly	30 arc	Gridded Population of the World data collection (GPW v4)	CIESIN-Columbia University (2017)
Longitude and latitude	Lon; Lat	Socioeconomic and coordinate variables	Not change by time	0.25°x0.25°		
Month	month	Socioeconomic and coordinate variables	Not change by time	0.25°x0.25°		

Table S2. The model performance for different regions: Southern Rocky Mountain (SRM), Southern California (SCA), Northern California (NCA), Pacific Northwest (PNW), southeastern US (SEUS), and northeastern US (NEUS)

	SRM	SCA	NCA	PNW	SEUS	NEUS	Whole US
Gird scale							
RMSE (km^2)	3.52	2.96	1.84	2.11	1.66	0.38	2.04
Correlation (r)	0.68	0.56	0.54	0.63	0.63	0.52	0.64
IoA	0.77	0.63	0.65	0.71	0.60	0.57	0.71
Percentage of grids with r > 0.4 (%)	64.5	70.1	84.6	80.4	82.7	80.4	76.0
Regional scale							
RMSE (km^2)	192.64	102.89	43.74	120.03	147.97	17.79	306.94
Correlation (r)	0.95	0.93	0.93	0.97	0.89	0.88	0.96
IoA	0.96	0.93	0.96	0.98	0.92	0.91	0.97

Table S3. The mean correlations between gridded observation and prediction for different regions and the p-values from t-test

	Mean correlation (with SVD predictors)	**Mean correlation (without SVD predictors)**	**p-value**
WUS	0.56	0.42	$< 2.2\times10^{-16}$
EUS	0.58	0.51	$< 2.2\times10^{-16}$
SEUS	0.59	0.52	$< 2.2\times10^{-16}$

Table S4. Performance metrics comparing simulation with and without longitude and latitude.

	RMSE	**IoA**	**Percentage of grids with temporal correlation > 0.4**
Simulation with all the predictors	2.03	0.71	79%
Simulation removing longitude and latitude	2.15	0.64	72%